PCT/JP99/02954
25.06.99

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

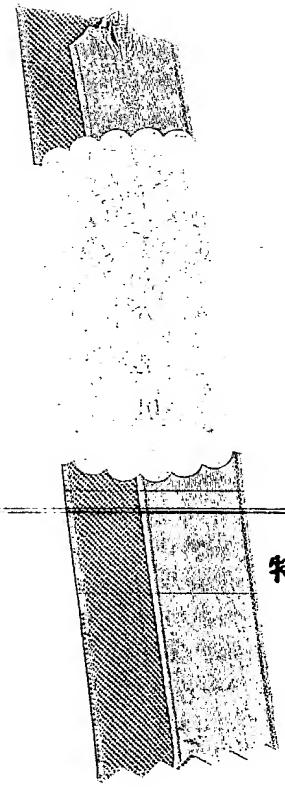
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９９年 ２月１９日

出願番号
Application Number: 平成１１年特許願第０４１１８６号

出願人
Applicant (s): 松下電器産業株式会社



PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

１９９９年 ７月１５日

特許庁長官
Commissioner,
Patent Office 伊佐山建志

出証番号 出証特平１１－３０５０２１４

【書類名】　特許願

【整理番号】　2015210011

【提出日】　平成11年 2月19日

【あて先】　特許庁長官殿

【国際特許分類】　G10L　3/00

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　脇田　由実

【特許出願人】

【識別番号】　000005821

【氏名又は名称】　松下電器産業株式会社

【代理人】

【識別番号】　100097445

【弁理士】

【氏名又は名称】　岩橋　文雄

【選任した代理人】

【識別番号】　100103355

【弁理士】

【氏名又は名称】　坂口　智康

【選任した代理人】

【識別番号】　100109667

【弁理士】

【氏名又は名称】　内藤　浩樹

【手数料の表示】

【予納台帳番号】　011305

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

出証特平１１－３０５０２１４

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9809938

【プルーフの要否】　不要

【書類名】　明細書

【発明の名称】　対訳フレーズ辞書作成装置

【特許請求の範囲】

【請求項１】　翻訳等の対象とされる文（以下、原言語文と称す）と、前記原言語文を異なる言語または文体などに翻訳した文（以下、目的言語文と称す）とを一対とした文データベース（以下、対訳コーパスという）において、意味的類似した単語を同クラスとみなして単語を分類し、同クラス内の単語に同コードを与えている表（以下、分類語彙表という）に基づき、対訳コーパスの各単語を左記分類コードに置き換える意味コード化部と、原言語文及び目的言語文における単語または上記単語クラスの隣接頻度を算出する隣接頻度算出部と、頻度の高い単語及び単語クラスを連結して意味的なまとまりを形成する部分文（以下、フレーズと呼ぶ）を抽出するフレーズ抽出部と、原言語及び目的言語のフレーズの関係を調べることで対応するフレーズを決定するする対訳フレーズ決定部と、決定された対訳フレーズを保管しておく対訳フレーズ辞書とから構成されることを特徴とする対訳フレーズ辞書作成装置。

【請求項２】　対訳コーパスにおいて、意味コード化部と、コーパスのパープレキシティー（文複雑度）を算出する文複雑度算出部と、上記意味コード化された対訳コーパスの文複雑度を用いてフレーズを抽出するフレーズ抽出部と、原言語及び目的言語のフレーズの関係を調べることで対応するフレーズを決定するする対訳フレーズ決定部と、決定された対訳フレーズを保管しておく対訳フレーズ辞書とから構成されることを特徴とする請求項１に記載の対訳フレーズ辞書作成装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、音声認識装置、言語翻訳装置、文体変換装置などで用いる対訳フレーズ辞書を作成するための対訳フレーズ辞書作成装置に関する。

【０００２】

【従来の技術】

以下、従来の技術を言語変換装置の１つである、入力音声を他言語に翻訳（以下通訳と呼ぶ）する通訳装置を例にして、通訳装置に用いる対訳フレーズ辞書として説明する。以下通訳装置における原体型文を原言語文、目的体型文を目的言語文と呼ぶ。

【０００３】

通訳装置は、音響信号として入力された発声文を単語テキスト列で表示された出力文に変換するための音声認識と、単語テキスト列で表示された文を入力し他言語文に翻訳する言語翻訳とを順次実行することで通訳を実現している。さらに上記言語翻訳部は、入力文の統語的または意味的構造を解析する言語解析部と、解析結果に基づいて他言語に変換する言語変換部と、翻訳結果から自然な出力文を生成する出力文生成部とから構成されている。

【０００４】

しかし、音声認識部が発声文の一部を誤認識した場合や、文にあいづちや言い直しなどが挿入されたり、一部の単語が省略されて話された場合など、文自体が統語的に不自然な場合は、音声認識結果を言語解析部に入力しても解析が失敗し、結果的に翻訳結果が出力されないという問題があった。

【０００５】

この問題に解決するために、従来の文法に則って言語解析を行うのではなく、従来の文法では解析できないような発声文も含めた発声文例から、対応する原言語文と目的言語文の対訳フレーズを抽出し、このフレーズ対をなるべく一般化した形で記述された対訳フレーズ辞書を作成し、この辞書を用いて言語解析と言語変換とを行う方法が提案されている。（たとえば、古瀬、隅田、飯田：情報処理学会論文誌Vol35,no3,1994-3）以下、文献の従来例を図５の対訳フレーズ辞書作成装置に沿って説明する。

【０００６】

通訳を行う前に、予め発声文対訳コーパスから対訳フレーズ辞書を作成する。

【０００７】

ここでは、一部の単語が誤ったり省略されたりすることを考慮し、発声文例をフレーズ毎に分割し、フレーズ内規則とフレーズ間の依存規則とを作成している

。

【０００８】

まず形態素解析部60で、原言語文と目的言語文との形態素解析を行ない、各文を形態素列に変換する。次にフレーズ決定部61で、原言語及び目的言語の形態素例をフレーズ単位に分割し、フレーズ内規則とフレーズ間の依存関係規則を作成する。この際のフレーズ単位は、意味的にまとまった単位であることに加えて、対訳において対応関係が明らかな部分文であることを考慮して人手で決定される。

【０００９】

たとえば、「部屋の予約をお願いしたいんですが」「I'd like to reserve a room」という対訳文例は、(a)「部屋の予約」「reserve a room」,(b)「をお願いしたいんですが」「I'd like to」という(a)(b)２つの対訳フレーズに分割され、「(a)を(b)する」「(b) to (a)」という依存関係が規則化される。上記対訳フレーズは対訳フレーズ辞書62に、フレーズ間の依存関係を対訳の形で表されたものはフレーズ間規則テーブル63に各々保管される。このような処理が対訳コーパスに含まれた全発声文分について行われる。このフレーズの分割と依存関係は、文の意味的情報やどの程度文法的に崩れていないかの度合いなどのファクターから決定されるため、自動的に各文について決定することが難しく、従来は人手で決定されている。

【００１０】

【発明が解決しようとする課題】

しかしながら、上記対訳フレーズ辞書やフレーズ間規則は、発声文の意味的情報や文法的情報を自動的に解析できる手段がないために、人手で作成しなければならない。そのため、開発に時間がかかり、人手の作成基準の揺れが規則性能を歪ませるという問題点がある。たとえば、通訳装置の目標となるタスクを変更したり、原言語及び目的言語の言語種が変更になった場合は、一度構築した規則を適応できずにはじめから規則を作成しなければならず、開発効率が悪く手間がかかる。また、上記フレーズ辞書やフレーズ間規則は、対訳コーパスの対応関係を重視してフレーズ単位を決定しており、音声認識部が認識するのに適切なフレー

ズ単位であるかどうかの評価がなされているものではない。

【００１１】

音声認識にとって適切なフレーズかどうかを人手で判断しながらフレーズ単位を決めることは困難であり、決定されたフレーズを用いて認識した場合、認識率が確保できる保証がない、という課題を有していた。

【００１２】

本発明の目的は、以上の課題点を解決し、フレーズ辞書作成やフレーズ間規則を認識及び言語変換の際に適切なフレーズ単位であることを考慮しながら、なるべく自動的に抽出することで、効率よく品質の高い対訳フレーズ辞書を生成できる対訳フレーズ作成装置を提供することにある。

【００１３】

【課題を解決するための手段】

上述した課題を解決するために、請求項１記載の対訳フレーズ作成装置は、分類語彙表に基づき対訳コーパスの各単語を左記分類コードに置き換える意味コード化部と、原言語文及び目的言語文における単語または上記意味コードに基づく単語クラスの隣接頻度を算出する隣接頻度算出部と、頻度の高い単語及び単語クラスを連結してフレーズを抽出するフレーズ抽出部と、原言語及び目的言語のフレーズの関係を調べることで対応するフレーズを決定するする対訳フレーズ決定部と、決定された対訳フレーズを保管しておく対訳フレーズ辞書とから構成される対訳フレーズ辞書作成装置を提供するものである。

【００１４】

また、請求項２に記載の対訳フレーズ作成装置は、請求項１に記載の意味コード化部とコーパスの文複雑度を算出する文複雑度算出部と、上記意味コードによる単語クラス化された対訳コーパスの文複雑度を用いてフレーズを抽出するフレーズ抽出部と、原言語及び目的言語のフレーズの関係を調べることで対応するフレーズを決定するする対訳フレーズ決定部と、決定された対訳フレーズを保管しておく対訳フレーズ辞書とから構成されることを特徴とする対訳フレーズ辞書作成装置を提供するものである。

【００１５】

次に本発明の動作を説明する。

【００１６】

請求項１に記載の対訳フレーズ作成装置は、単語間の意味的類似性を用いて単語クラスを作成し、隣接する頻度の高い単語または左記単語クラス列を連結してフレーズを生成し、原言語文及び目的言語文を対応させながら対訳フレーズを決定することにより、コーパスから対訳フレーズを自動的に抽出することを可能とし、人手をなるべく用いずに、効率よく品質の高い対訳フレーズ辞書を生成できる。

【００１７】

また、請求項２に記載の対訳フレーズ作成装置は、対訳フレーズを決定する際に、意味コードによる単語クラス化された対訳コーパスの文複雑度を用いて決定することにより、コーパスから対訳フレーズを自動的に抽出することを可能とし、人手をなるべく用いずに、効率よく品質の高い対訳フレーズ辞書を生成できる。

【００１８】

また、文複雑度の尺度が、音声認識に適切なフレーズかどうかの尺度と密接に関係があるため、認識精度を保証しながら、自動的にフレーズ抽出することが可能となる。

【００１９】

【発明の実施の形態】

（実施の形態１）

以下、図面を参照して本発明の請求項１の実施例を説明する。本実施例では従来例同様、通訳装置に用いる対訳フレーズ辞書作成装置として説明する。図1は本発明の請求項１に係る一実施例である対訳フレーズ辞書作成装置のブロック図である。

【００２０】

実施例の対訳フレーズ辞書作成装置は、形態素解析部２で対訳コーパス１内の原言語文の形態素解析を行うことで原言語文のみ品詞タグが付与された形態素列に変換する。たとえば、図２の「部屋の予約をお願いしたいんですが」の発声文

例では、ステップ21のような品詞タグが原言語文に与えられる。

【００２１】

次に、意味コード化部４で、原言語文の形態素列において、各形態素と分類語彙表に書かれている単語とを比較し、分類語彙表で意味コードが与えられている単語と一致した形態素については、形態素名を意味コードに置きかえることで、入力形態素列を一部の形態素が意味コード化された形態素列に変換する。

【００２２】

この際に意味コード化される形態素には以下の条件を満たすものとする。

【００２３】

（条件）対訳単語辞書に登録されている単語で、対訳単語辞書の目的言語訳に相当する単語が、コーパス内の相当する目的言語対訳文に存在する。

【００２４】

図２の例では、対訳単語辞書に登録されておりしかも分類語彙表でコードが与えられている「部屋」と「予約」のみが意味コード化され、ステップ22のようにこれらの形態素を意味コードに置き換えた形態素列が作成される。

【００２５】

さらに、相当する目的言語対訳文内の単語名もステップ22のように意味コードに置き換える。

【００２６】

次に、上記の一部の形態素が意味コードに置き換えられたコーパスについて、原言語文、目的言語文別々に、各単語または意味コードの２連鎖出現頻度（以後bi-gramと呼ぶ）を算出する。算出式を（数1）に示す。

【００２７】

【数１】

$$\text{bi-gram}=\frac{(\text{単語（又は意味コード）}i\text{と単語（又は意味コード）}j\text{が隣接して出現した数})^2}{\text{単語（又は意味コード）}i\text{の全出現数}\times\text{単語（又は意味コード）}j\text{の全出現数}}$$

【００２８】

コーパス内の全原言語文及び目的言語文を対象にbi-gramを算出した後、フレ

ーズ抽出部５で、最も出現頻度の高かった２単語または品詞対を１つの単語とみなして連結し、再度bi-gramを算出する。

【００２９】

これにより、たとえば頻度高く隣接する「お」「願い」、「願い」「し」、「し」「ます」などの単語対が連結され、「お願いします」というフレーズ候補が形成される。目的言語では「I'd」「like」、「like」「to」の単語対が連結される。全原言語文及び目的言語文別々に、以上の連結とbi-gram算出とを、bi-gramの値が全て一定閾値を超えなくなるまで繰り返す。そして、連結された単語も含めた個々の単語をフレーズ候補として抽出する。

【００３０】

次にフレーズ決定部６で、原言語文と目的言語文対において、各フレームが同時に出現している頻度を算出する。ｉ番目の原言語フレーズをＪ[ｉ]、ｊ番目の目的言語フレーズをＥ[ｊ]とすると、フレーズＪ[ｉ]とＥ[ｊ]との共起頻度Ｋ[ｉ，ｊ]は、算出式を（数２）にて算出される。

【００３１】

【数２】

$$K[i,j] = \frac{\text{フレーズ J[i] とフレーズ E[j] とが、対訳文対に共起する数}}{\text{フレーズ [i] の出現数} \times \text{フレーズ E[j] の出現数}}$$

【００３２】

たとえば、図３の例では、原言語文の３つフレーズのうち、原言語フレーズの「お願いします」と目的言語フレーズの「I'd like to」との共起頻度は２/（２＋３）、「したいんですが」と目的言語フレーズの共起頻度は１/（１＋３）となる。

【００３３】

この頻度が一定値以上のフレーズ対を対訳フレーズとし決定し、頻度と共にフレーズ番号を付けて対訳フレーズ辞書に登録する。さらに、対訳フレーズとして決定されなかったフレーズ候補の中で、既に品詞化されている単語は、それ単独

で対訳フレーズとして登録する。それ以外の部分は、対訳対の中で各々の単語列どうしを一対としてフレーズ辞書に登録する。

【００３４】

たとえば、図３の例では、ステップ31のようにフレーズ辞書に登録される。

【００３５】

このようにして、フレーズ登録を行なった後、一文に共起するフレーズ番号を記録し、フレーズ番号対としてフレーズ間言語規則に登録する。図３の例ではステップ32となる。

【００３６】

また、上記フレーズ番号フレーズbi-gramを求め、これもフレーズ間言語規則に記録する。

【００３７】

以上の実施例では、原言語文及び目的言語文各々における単語または意味コードの隣接頻度と、対訳における頻度の高い単語列または意味コード列の共起関係を用いて自動的に対訳フレーズとフレーズ間規則を決定し、この対訳フレーズ規則を用いて言語または文体変換とを行うことにより、コーパスから対訳フレーズを自動的に抽出することを可能とし、人手をなるべく用いずに、効率よく品質の高い対訳フレーズ辞書を生成できる。

【００３８】

（実施の形態２）

次に請求項２における実施例を説明する。本実施例も、先の実施例同様、異なる言語間の変換を行う通訳装置用対訳フレーズ辞書作成装置として説明する。図４は本発明の請求項２に係る一実施例である対訳フレーズ辞書作成装置のブロック図である。

【００３９】

実施例の通訳装置は、まず通訳する前に、先の実施例同様、形態素解析後、意味コード変換部で一部の形態素を意味コードに変換した対訳コーパスを作成する。

【００４０】

さらに、フレーズ抽出部で、原言語文、目的言語文別々に、各単語または意味コードのbi-gramを算出する。算出式は（数1）と同様である。

【００４１】

さらに、bi-gramの値が全て一定閾値を超えなくなるまで、先の実施例と同等に、処理を繰り返す。

【００４２】

そして、連結された単語も含めた個々の単語をフレーズ候補として抽出する。

【００４３】

次に、文複雑度算出部で、上記処理にて抽出されたフレーズ各々について、それがフレーズとなった場合と、フレーズとして連結されない場合との文複雑度を算出し比較する。文複雑度は（数３）で算出されるものである。

【００４４】

【数３】

$$\text{文複雑度 } F = 2H(L)$$

$$H(L) = -\sum_{1}^{M} P(Wi|Wi-1)\log P(Wi|Wi-1) / M$$

P(Wi|Wi-1)：i-1番目の形態素がWi-1であった時にi番目の形態素がWiである確率

M：全コーパスにおける2単語連鎖の種類数

【００４５】

比較した結果、フレーズ抽出部でフレーズにすることで文複雑度が増加するフレーズについては、フレーズ候補から除去する。

【００４６】

上記処理でフレーズ候補に残ったフレーズを対象に、先の実施例と同条件でフレーズを決定し、対訳フレーズ辞書とフレーズ間規則を決定する。

【００４７】

以上の実施例では、対訳フレーズを決定する際に、意味コードによる単語クラス化された対訳コーパスの文複雑度を用いて決定することにより、コーパスから対訳フレーズを自動的に抽出することを可能とし、人手をなるべく用いずに、効

率よく品質の高い対訳フレーズ辞書を生成できる。

【００４８】

また、文複雑度の尺度が、音声認識に適切なフレーズかどうかの尺度と密接に関係があるため、認識精度を保証しながら、自動的にフレーズ抽出することが可能となる。

【００４９】

【発明の効果】

以上詳述したように、請求項１に記載の対訳フレーズ作成装置は、単語間の意味的類似性を用いて単語クラスを作成し、隣接する頻度の高い単語または左記単語クラス列を連結してフレーズを生成し、原言語文及び目的言語文を対応させながら対訳フレーズを決定することにより、コーパスから対訳フレーズを自動的に抽出することを可能とし、人手をなるべく用いずに、効率よく品質の高い対訳フレーズ辞書を生成できる。

【００５０】

また、請求項２に記載の対訳フレーズ作成装置は、対訳フレーズを決定する際に、意味コードによる単語クラス化された対訳コーパスの文複雑度を用いて決定することにより、コーパスから対訳フレーズを自動的に抽出することを可能とし、人手をなるべく用いずに、効率よく品質の高い対訳フレーズ辞書を生成できる。

【００５１】

また、文複雑度の尺度が、音声認識に適切なフレーズかどうかの尺度と密接に関係があるため、認識精度を保証しながら、自動的にフレーズ抽出することが可能となる。

【図面の簡単な説明】

【図１】

本発明の実施の形態の構成を示す図

【図２】

本発明の実施の形態の動作を説明する図

【図３】

本発明の実施の形態の動作を説明する図

【図４】

本発明の実施の形態の構成を示す図

【図５】

従来の対話フレーズ生成蔵置を示す図

【符号の説明】

１　対訳コーパス

２　形態素解析部

３　分類語彙表

４　意味コード化部

５，４１　フレーズ抽出部

６，４２　フレーズ決定部

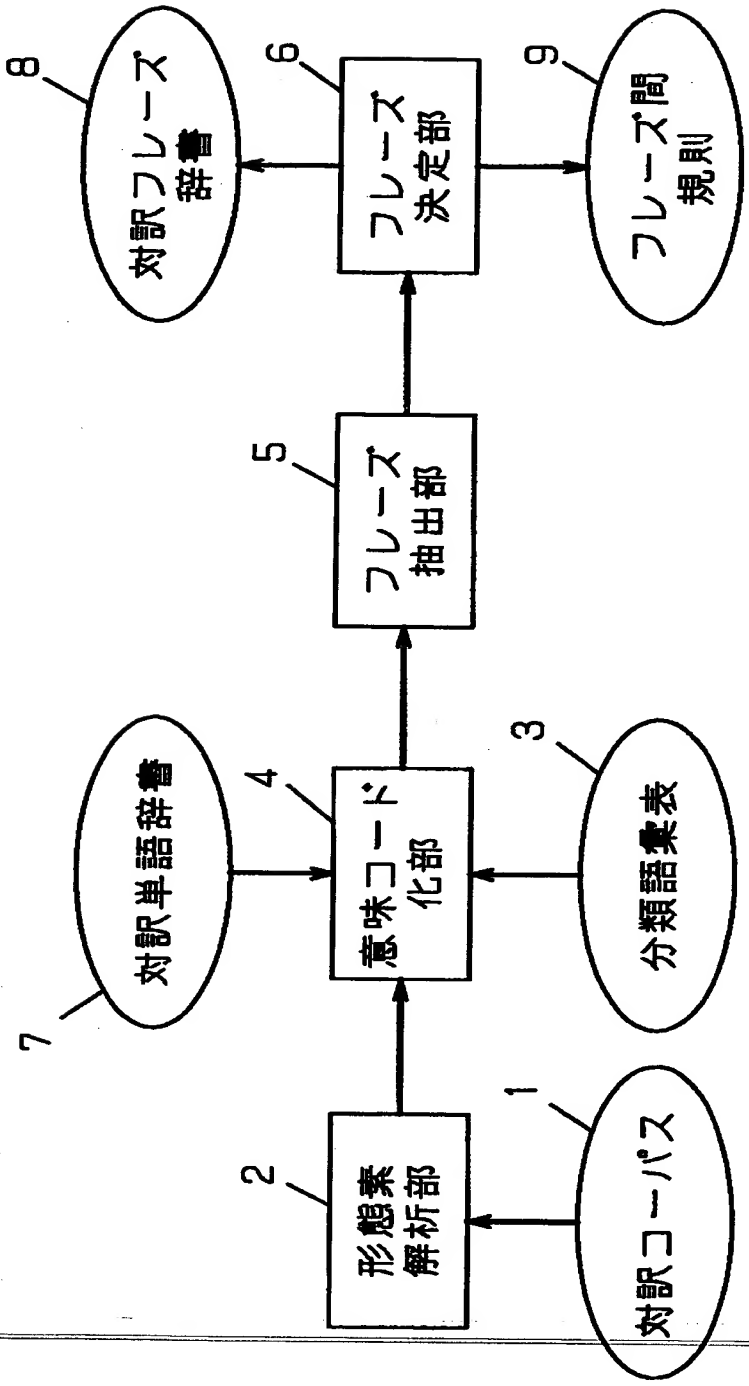
７　対訳単語辞書

８　フレーズ間規則

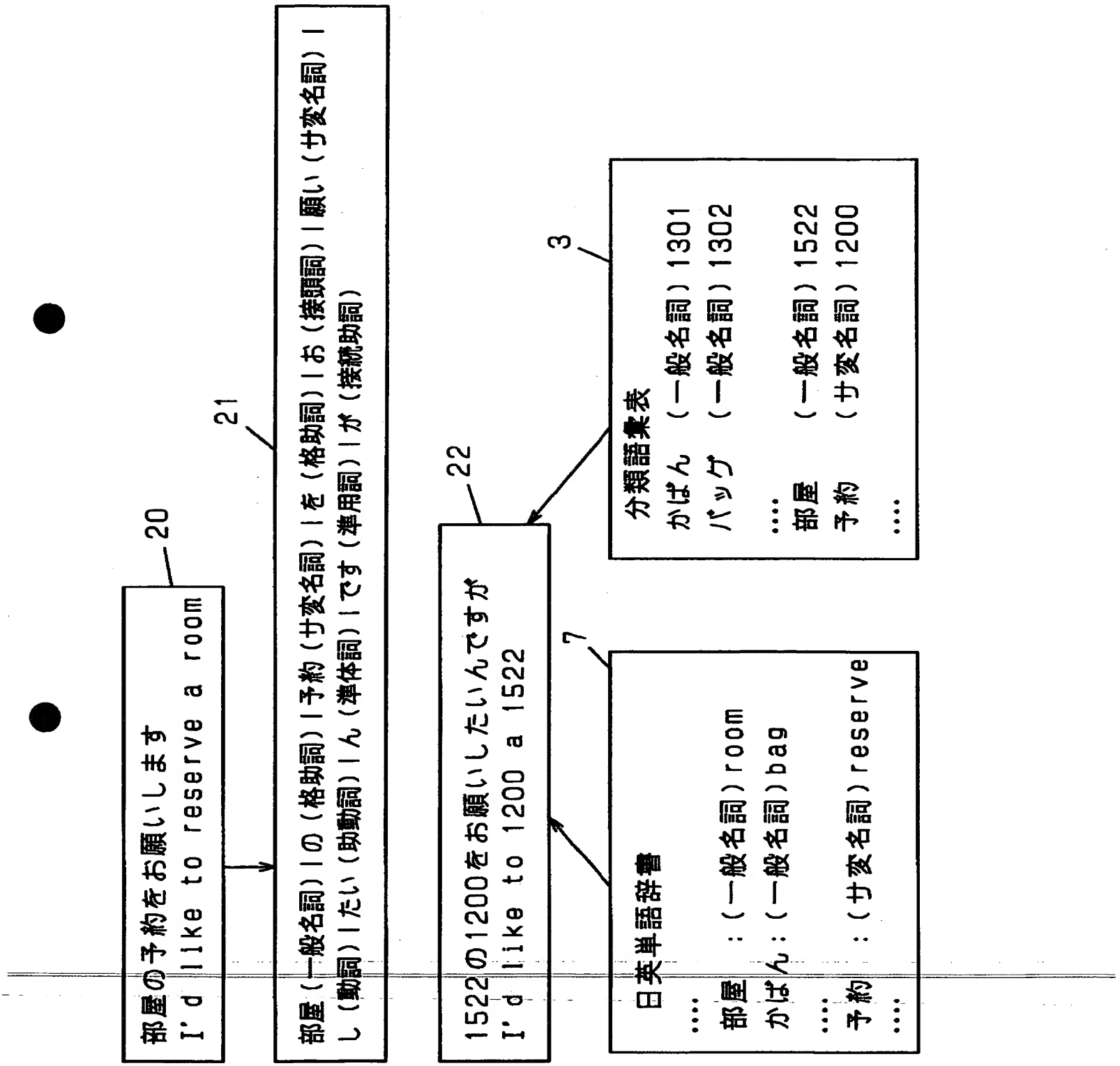
９　対訳フレーズ辞書

【書類名】　　図面

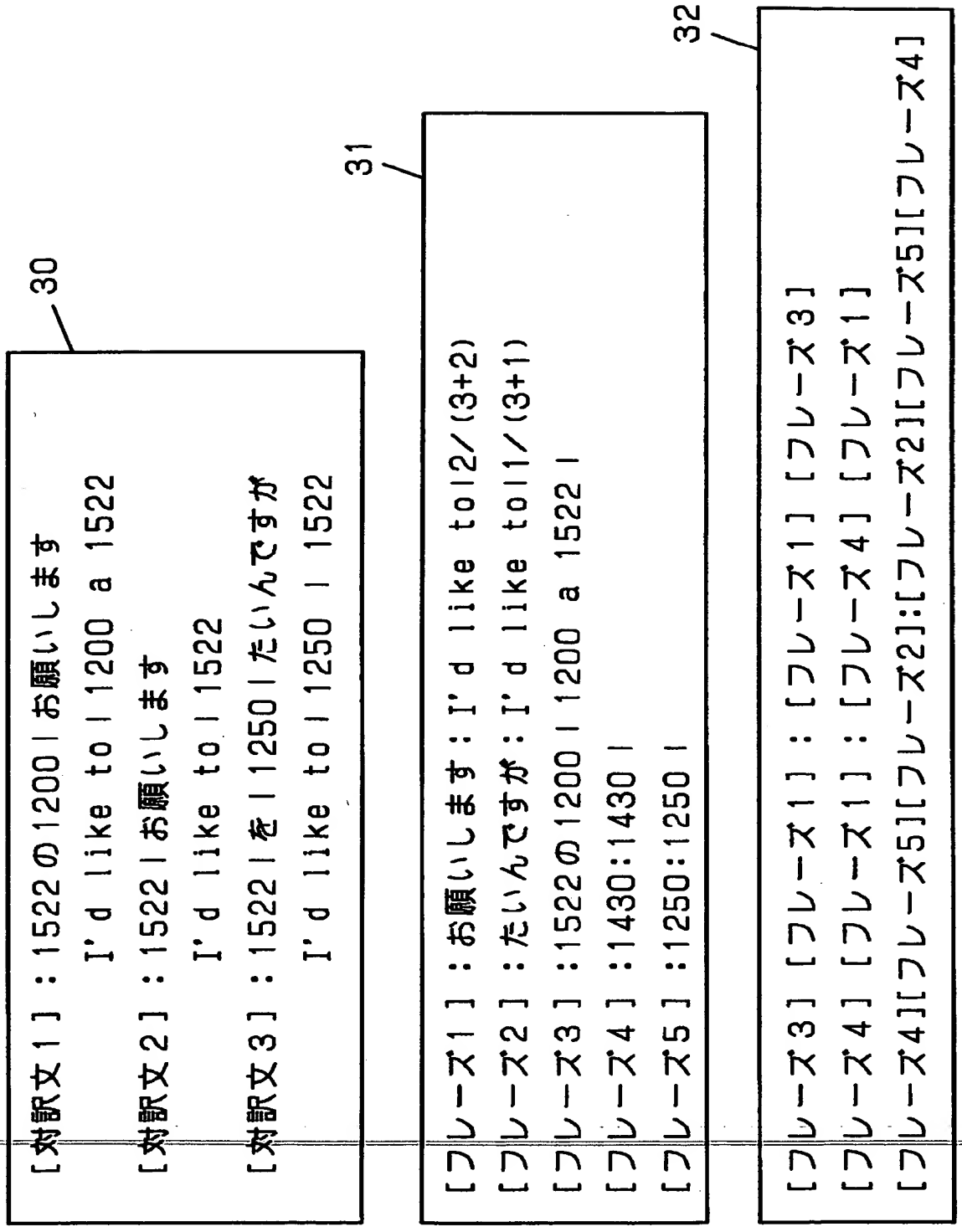
【図１】


対訳コーパス
1
形態素解析部
2
分類語彙表
3
意味コード化部
4
対訳単語辞書
7
フレーズ抽出部
5
フレーズ決定部
6
対訳フレーズ辞書
8
フレーズ間規則
9

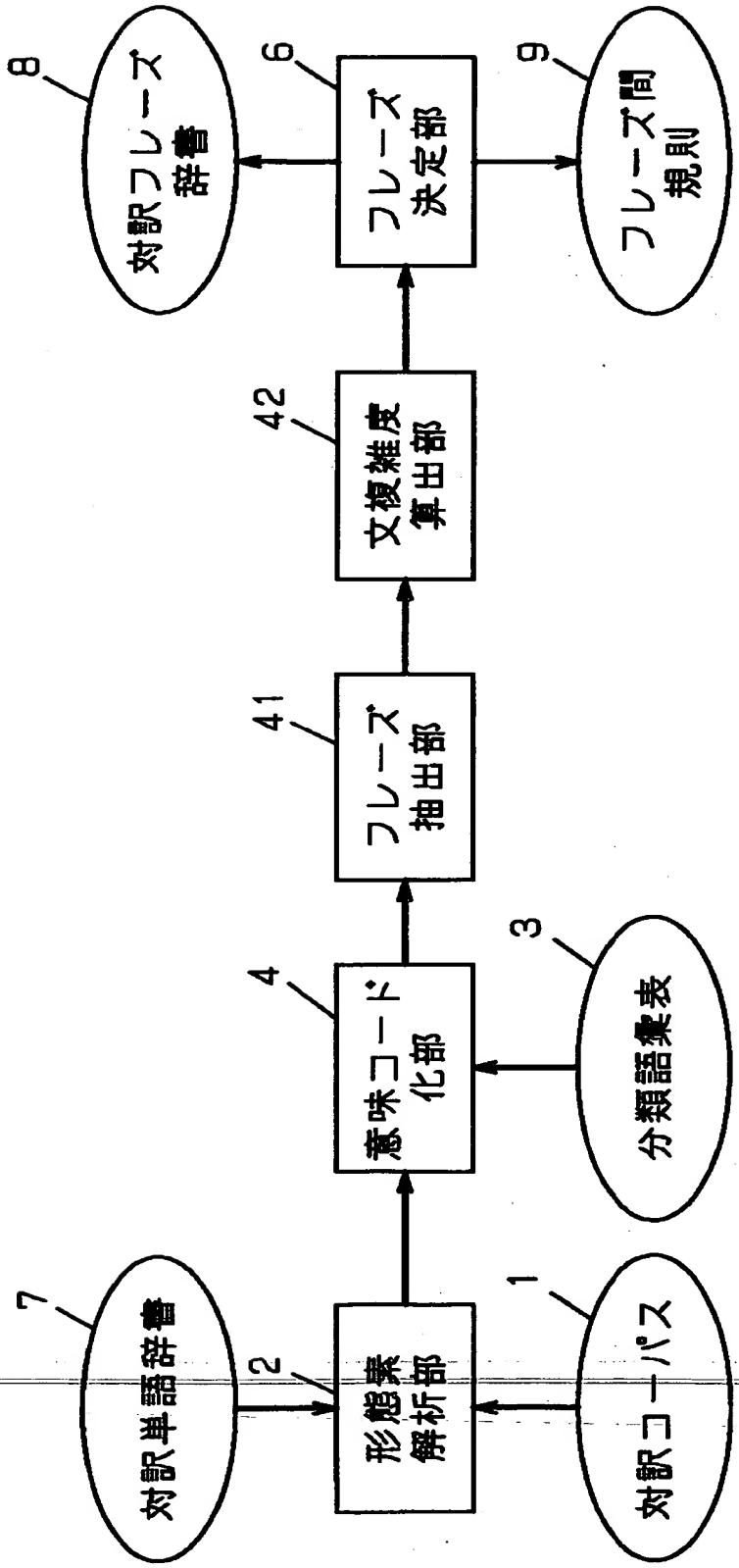
【図２】


20
部屋の予約をお願いします
I'd like to reserve a room
21
部屋（一般名詞）|の（格助詞）|予約（サ変名詞）|を（格助詞）|お（接頭詞）|願い（サ変名詞）|し（動詞）|たい（助動詞）|ん（準体詞）|です（準用詞）|が（接続助詞）
22
1522の1200をお願いしたいんですが
I'd like to 1200 a 1522
3
分類語彙表
かばん （一般名詞）1301
バッグ （一般名詞）1302
....
部屋 （一般名詞）1522
予約 （サ変名詞）1200
....
7
日英単語辞書
....
部屋 :（一般名詞）room
かばん :（一般名詞）bag
....
予約 :（サ変名詞）reserve
....

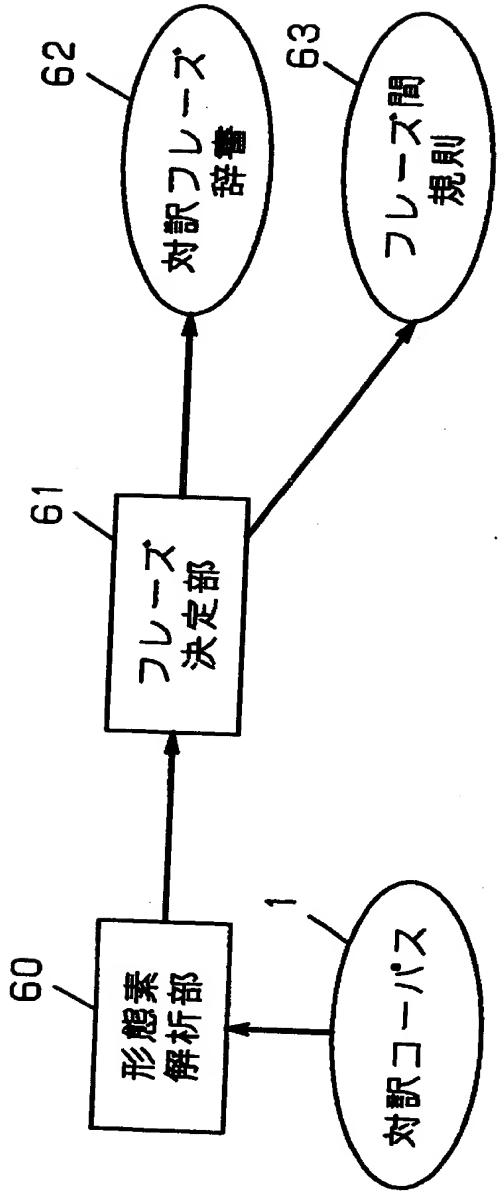
【図３】


30
[対訳文１]：1522の1200にお願いします
I'd like to 1200 a 1522
[対訳文２]：1522にお願いします
I'd like to 1522
[対訳文３]：1522に1250にたいんですが
I'd like to 1250 1522
31
[フレーズ１]：お願いします：I'd like to 2/(3+2)
[フレーズ２]：たいんですが：I'd like to 1/(3+1)
[フレーズ３]：1522の1200 1200 a 1522
[フレーズ４]：1430:1430
[フレーズ５]：1250:1250
32
[フレーズ３][フレーズ１]：[フレーズ１][フレーズ３]
[フレーズ４][フレーズ１]：[フレーズ４][フレーズ１]
[フレーズ４][フレーズ５][フレーズ２]:[フレーズ５][フレーズ４]

【図４】


対訳コーパス
1
形態素解析部
2
対訳単語辞書
7
意味コード化部
4
分類語彙表
3
フレーズ抽出部
41
文複雑度算出部
42
フレーズ決定部
6
対訳フレーズ辞書
8
フレーズ間規則
9

【図５】



【書類名】　要約書

【要約】

【課題】　人手で作成することなく、また開発時間を削減し、人手の作成基準の揺れによる規則性能の歪みを減少させる。また度構築した規則の改変が容易にできるようにする。

【解決手段】　隣接する頻度の高い単語または各単語の意味コード列を連結してフレーズを生成し、原言語文及び目的言語文を対応させながら対訳フレーズを決定することにより、コーパスから対訳フレーズを自動的に抽出することを可能とし、人手をなるべく用いずに、効率よく品質の高い対訳フレーズ辞書を生成できる。

【選択図】　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００５８２１〕

１．変更年月日　　１９９０年　８月２８日
［変更理由］　新規登録
住　所　　大阪府門真市大字門真１００６番地
氏　名　　松下電器産業株式会社